Evaluation of outreach interventions for under 16 year olds

Tools and guidance for higher education providers

Neil Harrison, Katy Vigurs, Julian Crockford, Colin McCaig, Ruth Squire and Lewis Clark

UWE Bristol University of the West of England

The University Of Sheffield.

Sheffield Hallam University

1. Overview

During 2017-18, OFFA commissioned research that aimed to understand the nature of outreach activities for under 16 year olds (which were funded through access and participation investment) and how these were evaluated.

This document, developed from the research, is intended to act as a resource for pre-16 outreach practitioners and evaluators, drawing both on the data collected by this project and the wider literature around evaluation and outreach. It seeks to recognise the complexity of pre-16 outreach work and eschews a prescriptive approach in favour of establishing important principles and actions that are likely to underpin good practice.

Our discussion is broadly positioned within a 'social realist' worldview (Archer, 2008; Pawson, 2013) that seeks to understand the fuzzy nature of the cause-and-effect relationships that exist within complex social fields, where individuals construct their own realities in reference to those around them. There is a particular focus on epistemology – the pathways to creating dependable, if contingent, knowledge – as a vehicle for making meaning from data that is usually incomplete, compromised or mediated through young people's emergent constructions of their worlds. Fundamentally, outreach is predicated on the ability of practitioners to influence young people in a planned way, albeit that the plan will not always work for every young person in every cohort.

An important element in this epistemology is that it is not concerned with finding single 'solutions' that exist outside time and context. Rather, it is concerned with understanding how young people are influenced by their life experiences – not 'what works', but what works in a given context and, importantly, why. It is only through understanding the latter element that practices can become robustly effective in the long-term and potentially transferable to other contexts. This is particularly appropriate to pre-16 outreach work due to the lengthy time lag between activity and application to higher education (HE).

2. Recommendations

The principal purpose of this project was to assist the Office for Students (OfS) with their policy development around the evaluation of pre-16 outreach work. However, we are also making three practice-focused recommendations to higher education providers (HEPs), which form the framing for this document:

1. We recommend that HEPs benchmark their evaluation practices against their peers with a similar organisational mission and profile of expenditure on access. We have developed a simple **self-assessment tool** to help HEPs to judge whether their practices are weakly- or well-developed, based on our findings among HEPs and third

sector organisations. This evaluation self-assessment tool has been provided to the OfS for further development and piloting.

2. We recommend that HEPs should seek to extend or enhance their evaluation practices to work towards the elements that characterise well-developed practice. We have developed a **development tool** to suggest small changes that HEPs can make to improve the rigour of their evaluations.

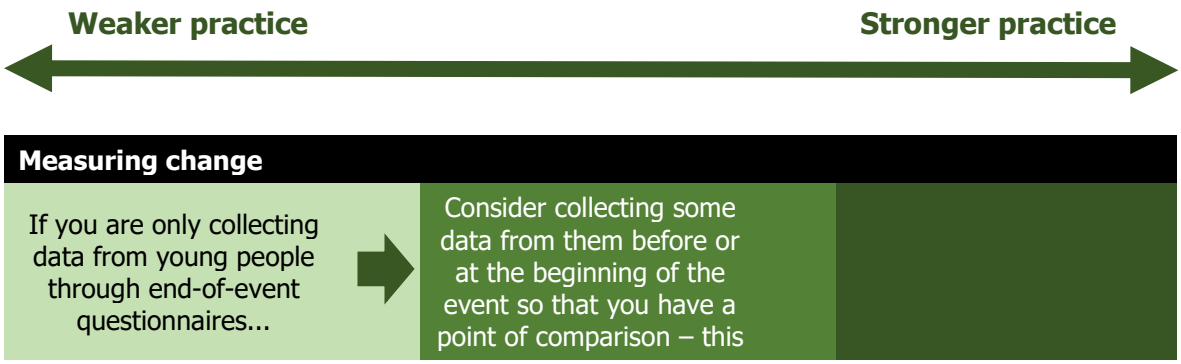
3. We recommend that HEPs should extend their understanding of the epistemology of evaluation to allow more robust claims to knowledge to be made. We have developed some **guidance and tools for providers to develop evaluation** to raise awareness of some of the key issues and to help HEPs to devise approaches to overcome these challenges. In particular, we recommend that HEPs consider adopting a 'theory of change' approach to planning and evaluating their activities.

These tools are intended to be read alongside Crawford et al. (2017) covering standards of evidence, by providing guidance to HEPs around the **standards of practice** that might generate stronger forms of evidence.

3. Development tool

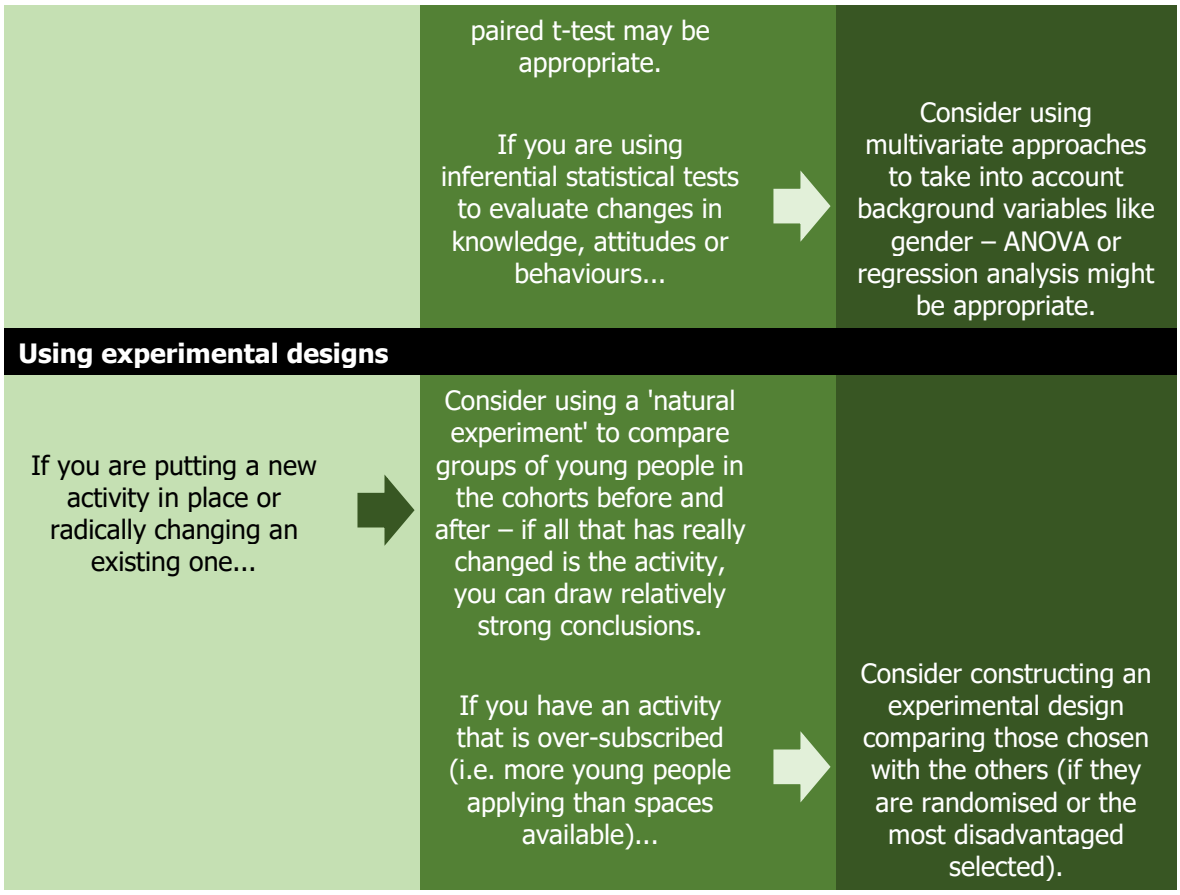
The purpose of this tool is to suggest practical ways in which HEPs can incrementally improve their practices within the framing provided in this report's overview. In some instances, this will involve upskilling staff or investing additional resources, but the suggestions are mainly resource-neutral or based around a conceptual shift.

This sort of tool is clearly not able to take account of where HEPs are on their journey towards stronger evaluation practice, so the suggestions will be more or less relevant to individual HEPs – they will probably be most useful for those HEPs with the least well-developed practice. Similarly, it is not able to cover all forms of pre-16 outreach activity, so we have focused on some general suggestions that will helpfully underpin high-quality evaluation regardless of the specific activity being evaluated.



	can be limited to a single question.	
	If you are collecting before and after data from young people...	Consider collecting data again some period later (maybe three to six months) to see whether any changes in knowledge, attitude or behaviour have remained.
Individualised data collection		
If you are currently using anonymous pre-post questionnaires to measure changes due to an activity...	Consider using identifiers/names to link data and look at individual change. This will enable you to determine effects more rigorously and see what proportion have changed.	
	If you are using linked individual data within pre-post questionnaires...	Consider looking at subgroups (e.g. by gender or ability) within the group to see whether the activity has been more effective for some types of young people than others.
Psychological and sociological concepts		
If you are interested in changes in young people's attitudes...	Consider tying your evaluation to well-established psychological or sociological constructs such as self-efficacy or social capital.	
	If you are using psychological or sociological constructs...	Consider using pre-existing inventories from the research literature as these are likely to have been validated – if none exists, ensure you cognitively test your own.
Collecting data from teachers and parents		
If you are currently relying on gathering evaluative data from young people...	Consider triangulating the self-report data by gathering data from the adults working with the targeted young people, including parents and teachers.	

	If you are currently collecting informal feedback from teachers, parents or other adults working with the targeted young people...	Consider using short telephone interviews – many will prefer this (response rates will be stronger) and you will collect richer data in a more robust way than using questionnaires.
Improving qualitative data collection		
If you are only collecting data from young people through questionnaires...	Consider undertaking focus groups or group interviews with a sample after a period of time has elapsed – this will give them the opportunity to reflect on their experiences.	
	If you are doing group interviews or focus groups with young people...	Consider taking steps to ensure that you have a balanced sample of young people involved and consider using an 'authentic task' exercise to provide additional observational data.
Exploring aspirations		
If the focus of your evaluation is on 'raising aspirations' for HE or similar...	Consider expanding your questions to take in the expectations of the young person, as research suggests these have stronger predictive power.	
	If you are already asking young people about their expectations around higher education...	Consider broadening out the questions to take in what they think their parents and teacher expect, as research suggests that these all have a strong correlation with future behaviour.
Inferential statistical analysis		
If you are using descriptive statistics (e.g. simple percentages) to measure changes in knowledge, attitudes or behaviours...	Consider using inferential statistical testing to determine whether the changes can safely be ascribed to the activity rather than chance – the	



4. Guidance for providers to develop evaluation

This section aims to provide some contextualised guidance to enable HEPs to engage more critically with key issues in evaluating pre-16 outreach. It is aimed primarily at those HEPs with moderately well-developed evaluation practice, although it is hoped that others will find elements useful too. As such, this section does not provide definitive instructions for how to evaluate particular activities, but rather raises questions that evaluators will need to consider in their own context.

4.1 Using evaluation methodologies

An interesting feature of the HEP survey data was the limited reported use of formal evaluation methodologies. Around one-third of institutions were using some form of pre-post design questionnaires, but these tended to be short-term investigations of immediate change in self-report data. Established evaluation methodologies offer the advantage of a rigorous epistemological foundation that has been developed and honed over multiple previous studies. In addition to logic chains and theory of change approaches, which are widely used in evaluation across many different fields, HEPs may consider exploring other methodologies.

Three common methodologies are briefly outlined below. These offer contrasting approaches that will be more or less applicable in different contexts – this list is not

intended to be exhaustive, although these enjoy widespread use within educational and other forms of social research:

- **Kirkpatrick model** (Kirkpatrick and Kirkpatrick, 2005, 2007) – based around four levels of evaluation, focusing on: (1) immediate reaction and satisfaction, (2) measured change in knowledge and/or attitudes, (3) sustained behavioural change, and (4) improved long-term outcomes. One tenet of the model is that evaluation practice tends to neglect levels 3 and 4, whereas activity planning should begin with these evaluations in mind.

- **Realist evaluation** (Pawson, 2006, 2013) – based on the premise that any activity is an embodied theory of change, the integrity and validity of which can be interrogated through different forms of data. Focuses on a conceptual model that considers: (1) the context, (2) the mechanisms for change, and (3) the desired outcomes. There is a particular focus on understanding why activities are effective as a means to understanding – and therefore replicating and enhancing – the causal mechanisms.

- **Evaluative case study** (Yin, 2018) – based on the assumption that the effectiveness of an activity is closely entwined with its physical and human context and that a holistic approach is required. With a strong emphasis on triangulation of method and perspective, the evaluation is constructed around 'theoretical propositions' that are tested and refined through the data collected.

All three of these methodologies are predicated to some extent on a theory of change approach, albeit described in differing terminology. In addition, the University of Bath has developed the 'Network for Evaluating and Researching University Participation Interventions' (NERUPI) model[1] for evaluating outreach work and you may wish to explore what this has to offer.

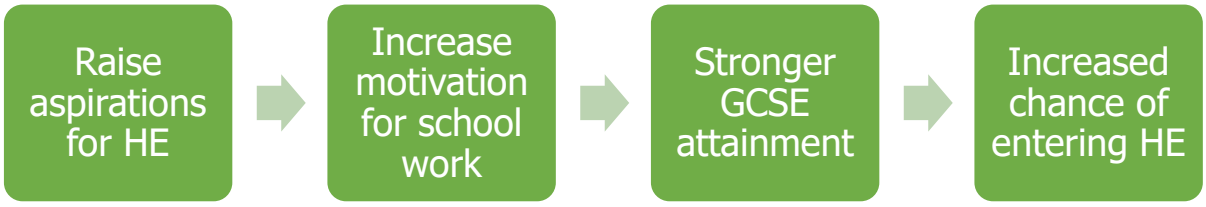
4.2 Developing theory of change and logic chains

A key element in robust evaluation can be the articulation and testing of a clear 'theory of change' for a given activity – this was common among the third sector organisations interviewed in this project, but much less so for HEPs. It requires organisations to explore in detail the mechanisms that underpin its activity, as described by Harries et al. (2014, p.5):

> 'Theory of change is actually a very simple concept. Throughout our work and personal lives we have aims, objectives and ideas about how to achieve our goals, but we rarely take the time to think these through, articulate and scrutinise them. All a theory of change process does is to make these assumptions explicit and therefore more testable'.

[1] See www.nerupi.co.uk.

In other words, it lays out the **intermediate steps** between the young person's current state and the ultimate goal – in most cases for pre-16 outreach interventions, this will be entry into HE. A particular feature of pre-16 outreach is the length of time between intervention and ultimate outcome, which increases the importance of theorising change in a clear way. An example of a commonly-expressed theory of change for pre-16 outreach work is presented below:



In this instance, the only link in the logic chain that is well-evidenced in the research literature is the final one (Crawford, 2014). As noted in Section 5 of the main report for our research ('Understanding the evaluation of access and participation outreach interventions for under 16 year olds'), the link between aspirations for HE and motivation for school work is questionable, while there is even limited evidence for a direct link between motivation and attainment (Cummings et al., 2012; Gorard and See, 2013).

Under a theory of change approach, the purpose of evaluation is to evidence the links in the logic chain and especially those where the validity of the assumptions made is most questionable. Clearly every activity will have its own chain, which may have multiple chains within it, and therefore its own evidential challenges. Efforts should be focused on those elements of the logic chain that are felt to have the weakest underpinning evidence, either from the global research literature or within the specific setting.

This constant questioning process has two main advantages over more traditional approaches to evaluation. Firstly, it allows the long-term aim (participation in HE) to be held in mind over a long period and over multiple intermediate steps that underpin behavioural change. Secondly, it can provide for a clear articulation of the vital sub-elements in an outreach activity, rather than always focusing on the whole so that key details can get lost.

You may wish to explore the following links to materials about developing and using a theory of change approach:

- www.nesta.org.uk/sites/default/files/theory_of_change_guidance_for_applicants_.pdf
- www.open.edu/openlearncreate/course/view.php?id=2214
- www.devinfolive.info/impact_evaluation/img/downloads/Theory_of_Change_ENG.pdf
- www.thinknpc.org/publications/theory-of-change

4.3 Identifying and testing intermediate steps

As noted above, a strong advantage of using a theory of change approach to evaluation is that it enables the identification of key intermediate steps between an outreach activity and future behaviour (i.e. HE application).

The nature of these intermediate steps is necessarily dependent on the activity or programme being evaluated, but their clear articulation should provide alternative opportunities for measuring and understanding the outcomes. These intermediate steps might be derived from a consideration of existing activities, the research literature or social theory, and might make use of established concepts from sociology (e.g. cultural capital), psychology (e.g. self-efficacy) or other social science disciplines.

The use of established concepts offers the additional advantage that there are likely to be pre-existing evaluation tools and metrics that can be adopted. Those developed in the research literature are likely to have been cognitively tested and validated across multiple populations, although it should be remembered that their applicability to a new context needs to be established; for example, a 'crisis of replicability' exists in the discipline of psychology due to the overuse of US undergraduates as study participants in devising new concepts and metrics.

By shortening the timescales for measured outcomes, the use of intermediate steps gives the opportunity to make stronger causal claims about outreach activities than those provided by very long-term perspectives where multiple confounding factors make it harder to disentangle influences on young people's decision-making (Harrison and Waller, 2017).

4.4 Making safe inferences from self-report data

As touched on in Section 5 of the main report ('Understanding the evaluation of access and participation outreach interventions for under 16 year olds'), there are significant issues with the reliability and validity of self-report data, especially from younger age groups. These can take a number of forms and it is useful for practitioners to bear these mind, alongside more general good practice in questionnaire design and implementation:

- **Logistics** – young people may not engage fully with completing questionnaires or may give inaccurate responses, especially if there is insufficient time or if the questionnaires are poorly designed. This may be more likely among particular disadvantaged groups such as those with dyslexia or English as an additional language.

- **Placebo effect** – participants will tend to alter their behaviour when they know they are being researched. Participation in an outreach activity is a very clear signal to a young person that they are expected to valorise extended education or have an increased interest in a subject area. Their responses to questionnaires are likely to reflect these expectations to some extent, regardless of the content or pedagogy embodied in the activity, and this will lead to an overestimate of effect.

- **Priming effect** – participants in an activity will tend to rate it more positively the nearer in time that they are asked about it, especially if they have enjoyed it and if the most enjoyable elements came last. Self-report questionnaires completed immediately at the end of an event are likely to overestimate its effect, particularly where participants are asked to project this impact into the future: e.g. 'Are you more likely to apply to HE?'

- **Dunning-Kruger effect** – people who are unknowledgeable or unskilled in a particular field tend to overestimate their level of knowledge of skill due to the lack of a reference point. This might occur where a young person feels that their knowledge has fallen after an event as it made them realise that they knew less than they thought. This might happen particularly with intellectually challenging activities.

- **Social desirability bias** – participants will tend to want to give answers that they think are expected of them or that will please the evaluator. This effect is usually stronger in interviews and non-anonymous questionnaires, but it impacts on all data collection and tends to exaggerate effect sizes.

- **Self-selection bias** – certain categories of people (e.g. women) tend to be more likely to engage with questionnaires fully, so findings can become badly skewed if only certain people provide data. With outreach activities, the risk is that participants who have benefited from the activity will be more likely to provide data.

- **Gender and ethnicity** – there is some tendency for women and people from minority ethnic communities to self-report personal attributes like confidence or self-esteem at a lower level than their peers. This may be particularly salient when comparing different subgroups.

These challenges do not invalidate the use of self-report data, but they do mean that its interpretation needs caution. There may be ways of mitigating the challenges, for example by collecting data after a delay or triangulating self-report data against other data (e.g. from teachers or school tests). More importantly, they mean that claims about effectiveness should be carefully constructed – e.g. by avoiding assertions of 'proof' or by comparing multiple years of data before drawing conclusions about the effectiveness of the activity.

A more general point about self-report data collected through the use of rating or Likert scales is the tempting assumption that this form of data is inherently more 'scientific' or accurate than qualitative data. In essence, a questionnaire is simply a very structured form of interview where the evaluator is (usually) absent. The person responding is still providing a highly subjective assessment, but with the added disadvantage that they are unable to ask for clarification. For example, a questionnaire about self-confidence does not provide a direct measure of a young person's self-confidence, but rather what they wish to say about their own subjective assessment of their self-confidence – the measure is two steps removed from the phenomenon that it seeks to measure.

Of course, there are also advantages from the use of questionnaires to collect quantitative data; for example, the person responding may feel under less pressure to provide an 'expected' answer due to perceived or real anonymity. The main advantage they provide is the ability to quickly gather large numbers of subjective approximations, which is obviously extremely useful for statistical analysis. However, these approximations are not inherently superior to self-report data collected through qualitative methods – the latter can, of course, be readily turned into quantitative data.

4.5 Cognitive testing your questionnaires

When developing evaluation tools such as questionnaires, we need to understand how potential recipients understand them and to establish that their responses are meaningful to the data we want to collect – i.e. that they have internal validity. Issues around respondents' comprehension, memory and willingness to respond honestly or completely can be identified and addressed.

Cognitive testing is often done via a focus group with a pilot group of respondents, either as they respond to your evaluation questions or shortly afterwards, and can involve asking participants to describe how they understood questions and what informed their response. This process is particularly important with young people, who might have a completely different understanding of seemingly straightforward questions from outreach practitioners.

You may wish to explore the following links to materials about cognitive testing:

- www.gov.scot/Resource/Doc/175356/0091403.pdf
- www.gesis.org/fileadmin/upload/SDMwiki/LenznerNeuertOtto_Cognitive_Pretesting.pdf

4.6 Using linked individual data

One tension in evaluation concerns whether or not data should be collected anonymously from participants. On the one hand, anonymity allows young people to feel more free to

express their opinions and to be less concerned about data security. This may lead to more accurate data in some situations.

On the other hand, anonymity makes it considerably harder for the evaluator to measure and assess changes resulting from an activity. Rather than focusing on how individuals have learned or been influenced, the evaluator is forced to look only at the cohort as a whole. While there appears to be some value in being able to demonstrate that XX% of a cohort held an opinion before an activity and YY% after, this is less helpful in building causal claims for effectiveness than it appears, as:

- To be valid, the response rates for the questionnaires need to be very high, particularly within a pre-post design. If they are not, then any apparent changes in the cohort could be simply due to different young people responding.

- Without measures of change for individuals, it is impossible to determine whether the activity is more or less successful for different demographic groups – or even potentially harmful for some. In other words, it precludes subgroup analyses that provide a richer picture of effectiveness.

One simple approach, used by most of the case study institutions, is to ask for young people's names to allow multiple data points to be connected, being clear that their data was not anonymous. However, there are other potential approaches that provide anonymity alongside the ability to match individuals. For example, code numbers could be used, proxies for matching by postcode, the use of a 'honest broker', two-sheet questionnaires and so on.

4.7 Developing experimental and quasi-experimental designs

Only one HEP reported that it was currently using an experimental design as part of its pre-16 outreach evaluation; in this case, a randomised controlled trial (RCT). This may be, in part, because the long timescales and social complexity involved do not lend themselves to the isolation of single causal effects. Indeed, in order to achieve the evidential value that is posited for experimental designs, there needs to be rigour in the application of the approach – a poorly-conducted experimental design is considerably less useful than a well-conducted pre-post design or rigorous qualitative study.

However, the use of intermediate steps (as described above) may make this more achievable if the following issues can be overcome:

- **Sample size.** The size of sample needed is a function of the effect size that the activity is expected to have – for small activities in a complex social field, experimental

and control groups of fewer than 50 are unlikely to provide sufficient power to identify significant effects.

- **Randomisation.** It is unlikely that there will be an opportunity for randomisation unless access to an activity in oversubscribed and the institution has the ability to select participants randomly. Even then, the numbers of individuals involved would have to be high in order to provide a control group that was reliably similar to the experimental group.

- **Controlling.** The concept of controlling extends beyond simply constructing a control group. A robust experiment would also have measures in place to ensure that, for example, there was no contamination between the groups (e.g. by young people talking about the activity to friends) and some means in place to account for the placebo effect.

Of course, it is possible to devise experimental designs that are not randomised controlled trials, but that still have high evidential value. For example, an alternative to randomisation is to purposively match members of the experimental and control groups across salient demographic and educational variables (e.g. gender, ethnicity or KS2 attainment) in order to achieve balanced groups. Another alternative is to use forms of natural experiment where, for example, the experimental and control groups are chosen from different points in time before and after a new activity is implemented.

The use of quasi-experimental approaches, where the participants in an activity are compared retrospectively to non-participants, was alluded to by institutions, but no firm examples were provided. This may be particularly appropriate with respect to analysing tracking datasets and other instances where the evaluator is seeking to explain outcomes for some young people within a wider population – e.g. those in a school who have been part of an outreach programme, compared to those that have not.

The principal challenge within quasi-experiments is defining a valid comparison group. The group should be as close as possible to the experimental group, with participation (or not) in the activity being the only meaningful difference. However, due to the targeting that is inherent in outreach work, this is very difficult to achieve – fundamentally, there is nearly always a meaningful reason why one young person was chosen to participate while similar ones were not, including perceived potential for HE, parental support, family commitments and so on.

References

Archer, M. (2008) *Realist Social Theory: The Morphogenetic Approach*, Cambridge: Cambridge University Press.

Crawford, C. (2014) *The Link between Secondary School Characteristics and University Participation and Outcomes*, London: Department for Education.

Crawford, C., S. Dytham and R. Naylor (2017) *The Evaluation of the Impact of Outreach: Proposed Standards of Evaluation Practice and Associated Guidance*, Bristol: Office for Fair Access.

Cummings, C., K. Laing, J. Law, J. McLaughlin, I. Papps, L. Todd and P. Woolner, P. (2012) *Can Changing Aspirations and Attitudes Impact on Educational Attainment? A Review of Interventions*, York: Joseph Rowntree Foundation.

Gorard, S. and B.H. See (2013) *Overcoming Disadvantage in Education*, Abingdon: Routledge.

Harries, E., L. Hodgson and J. Noble (2014) *Creating Your Theory of Change*, London: New Philanthropy Capital.

Harrison, N. and R. Waller (2017) Evaluating outreach activities: overcoming challenges through a realist 'small steps' approach, *Perspectives: Policy and Practice in Higher Education*, 21(2/3), 81-87.

Kirkpatrick, D. and J. Kirkpatrick (2005) *Transferring Learning to Behavior: Using the Four Levels to Improve Performance*, San Francisco: Berrett-Koehler.

Kirkpatrick, D. and J. Kirkpatrick (2007) *Implementing the Four Levels: A Practical Guide for Effective Evaluation of Training Programs*, San Francisco: Berrett-Koehler.

Pawson, R. (2006) *Evidence-Based Policy: A Realist Perspective*, London: Sage.

Pawson, R. (2013) *The Science of Evaluation: A Realist Manifesto*, London: Sage.

Yin, R. (2018) *Case Study Research and Applications: Design and Methods* (6th edition), London, Sage.